State of California Department of Consumer Affairs

OFFICE OF EXAMINATION RESOURCES
501 S Street, Suite 3
Sacramento, CA 95814

July 3, 2001

Mr. Thomas O'Connor
Executive Officer
Board of Psychology
1422 Howe Avenue # 22
Sacramento, CA 95825-3200

Dear Mr. O'Connor:

The Office of Examination Resources evaluated the oral examination program of the Board of Psychology according to professional testing standards. Although the oral examination program has served the needs of the Board of Psychology in the past, it is now time to ask questions about the validity of its licensing decisions. The information provided in the document attached to this letter identifies the problems and suggests that it might not be possible to offer an oral examination that will meet the Standards for Educational and Psychological Testing.

It is the recommendation of the Office of Examination Resources to discontinue the oral examination. If you have questions, I can be reached at 916.322.2703.

Norman R. Hertz, Ph.D., Chief
Office of Examination Resources

c: Denise Brown, Chief Deputy Director
Lynn Morris, Deputy Director, Board Relations

THE ORAL EXAMINATION FOR LICENSURE OF PSYCHOLOGISTS

INTRODUCTION

The Office of Examination Resources of the California Department of Consumer Affairs evaluated the oral examination of the Board of Psychology. The analysis, part of the board's ongoing review of the oral examination, is to determine whether the examination is fulfilling its intended purpose. To add or delete portions or even change the entire examination does not invalidate previous examination decisions. All regulatory agencies have the obligation to evaluate their examination programs even in the absence of challenges to the validity of the programs.

The board's oral examination has been developed to high standards and has been serving as a means of assessing minimum competence. The content for the examination is based upon job-related behaviors and associated knowledge defined in a practice analysis conducted by the American Association of State and Provincial Psychology Boards. Structured questions and responses were developed, behaviorally anchored rating scales were developed to use as the criteria for performance levels, and standardized test administration procedures were implemented. The decision to pass or fail the candidate is based upon an aggregate of their scores in response to the structured questions.

However, the oral examination is a "test" and must meet all the psychometric standards of the Standards for Educational and Psychological Testing.[1] Despite efforts by the Board of Psychology to improve the psychometric quality of the oral examination, questions remain about the validity of the pass/fail decisions. The oral examination may not be measuring a candidate's potential for safe practice because of problems in development, administration, and scoring. It is necessary to consider the oral examination in the light of its consequences of failing or passing candidates, not solely in relationship to the intent of the examination. The constructs being assessed are not testable in the context of a licensing examination.

There are several sources of error in an oral examination that adversely impact the standardization, and thus, the validity of the oral examination. One source of error has to do with content, and the others are procedures for examination development and administration. Granted, the content of the oral examination is linked to a practice analysis to control for error related to content. However, the content of the examination requires a second interpretation by examiners. The development of the vignette and the questions depends upon interpretation of the content of the practice analysis by psychologists in the development workshops. Regardless of the psychometric quality of the performance criteria, they have value only if they are applied consistently according to the level of candidates' performance. All candidates should have the same examination experience, not only in terms of administration, but also in terms of scoring.

---

[1] American Educational Research Association, American Psychological Association & National Council on Measurement in Education. (1999). Standards for Educational and Psychological Testing. Washington, DC: American Educational Research Association.

RELEVANT STANDARDS

The Standards for Educational and Psychological Testing are the standards by which the oral examination was evaluated. The most relevant standards to oral examination are:

Standard 2.10	*"When subjective judgment enters into test scoring, evidence should be provided on both inter-rater consistencies in scoring and within-examinee consistency over repeated measurements. A clear distinction should be made among reliability data based on (a) independent panels of raters scoring the same performances or products, (b) a single panel scoring successive performances or new products, and (c) independent panels scoring successive performances or new products." (p. 33-34)*
Standard 3.22	*"Procedures for scoring and, if relevant, scoring criteria should be presented by the test developer in sufficient detail and clarity to maximize the accuracy of scoring. Instructions for using rating scales or for deriving scores by coding, scaling, or classifying constructed responses should be clear." (p. 47)*
Standard 3.23	*"The process for selecting, training, and qualifying scorers should be documented by the test developer. The training materials, such as the scoring rubrics and examples of test takers' responses that illustrate the levels on the score scale, and the procedures for training scorers should result in a degree of agreement among scores that allows for the scores to be interpreted as originally intended by the test developer. Scorer reliability and potential drift over time in raters' scoring standards should be evaluated and reported by the person(s) responsible for conducting the training session." (p. 47-48)*
Standard 3.24	*"When scoring is done locally and requires scorer judgment, the test user is responsible for providing adequate training and instruction to the scorers for examining scorer agreement and accuracy. The test developer should document the expected level of scorer agreement and accuracy." (p. 48)*

VALIDITY ISSUES

While the general content of the oral examination is based on a practice analysis, there are several important issues related to the validity of the examination that are in question.

Constructs tested. The examiners are asked to evaluate candidate performance in various subject matter areas although the resultant scores are used to predict whether or not the candidate will practice safely once he or she is licensed. The critical question is one of

whether the candidate will harm the public once he or she is licensed. Perhaps the constructs assessed in the oral examination may not be testable in the context of a licensing examination.

Clinical content of the examination. Psychologists who developed the examination were instructed to maintain psychometric standards throughout the development process. However, the content of the clinical issues within vignettes, the structured questions, and the criteria responses depended largely upon the interpretation of the psychologists.

Examination administration procedures. The vignette developers design the content of clinical scenarios and criteria responses that they have determined to be appropriate for entry-level practice. However, examiners may interpret the content of the vignette and the criteria responses. Different sets of examiners produce vastly different ratings. Regardless of the psychometric quality of the performance criteria, these have value only if applied consistently.

Examiner training procedures. A major source of error in the oral examination is the lack of ability to calibrate the examiners' judgments. If the examiners are not calibrated to the same standard, candidates do not receive the same opportunity to be successful. Some error is controlled by linking the content of the examination to a practice analysis; however, other sources of error can be managed somewhat with extensive training, supervision, and feedback. Given the wide range of experience, training, and expectations of examinations, the training would have to calibrate the examiner to a shared definition of minimum competence. But, even under the best of conditions, there will be variability among and within the teams of examiners.

FINDINGS

The findings are based upon statistical analyses presented to the board after each oral examination cycle, formal discussions regarding validity of the examination by focus groups, and blind study of examiner ratings.

Correlation. Intraclass correlation coefficients have been used to provide an estimate of internal reliability of the examination. The coefficients have been quite high and thus indicate that the examiners were consistent. The percentage of agreement between examiners' decisions to pass or fail candidates has also been quite high. However, there are several indications within the same administration and across administrations that the validity of the examination is questionable.

Passing rates. Historically there have been differences between the passing rates in the northern and southern California sites. The examination is offered in the Los Angeles area on one day and then offered in the San Francisco area a week later. Minor modifications to the vignette, structured questions, and criteria responses are made prior to the San Francisco administration. Typically, the changes amount to modifications of the test data and MMPI-2 profile.

Over the last five administrations (January 1999, June 1999, January 2000, June 2000, and January 2001), passing rates for the two sites vary as little as 1% and as much as 11%. About half of the time, the passing rate of candidates who take the examination in Los Angeles is greater than the candidate pass rate in San Francisco. The same locale pass rate can also be reversed about half of the time, where candidates in San Francisco show the most successful pass rate.

The only real differences between the two sites are the examiners. The orientation, test administration procedures, and vignette are virtually identical. The only variables common to the sites are the examiners and the candidates. There is no discernible rationale for the difference in the qualifications of the candidates in regard to where they choose to sit for the examination. The assumption may be made that the only difference is the examiners and whether they are using different standards and scoring methods. The magnitude of the difference and the lack of consistency in the passing rate suggest that the examination process itself or the examiners' scoring procedures is flawed (see Standards 2.10 and 3.22).

Examiner training. Training for the oral examiners consists of a one-hour orientation on the morning of the examination. The training makes no attempt to calibrate the standards of the examiners so that they have a shared understanding of the expectations for minimum competence (see Standards 3.32 and 3.24). One of the primary principles of testing is standardization. That is, candidates must be offered the same examination in all aspects. Without training, the examiners are subjectively interpreting the meaning of the rating criteria. Without a systematic scoring process, candidates are held to different standards, higher or lower, depending upon their luck.

Reliability and validity. The Board of Psychology and the Office of Examination Resources conducted a study in October 1999 to examine the reliability of the examiners and the validity of their licensing decisions. For the study, 15 experienced examiners were convened as raters to listen to the tapes of four candidates who sat for the oral examination previously. The instructions to the raters were to follow the same procedures as they did when they served as examiners. The original scores for the four candidates were also available. The tapes were selected to represent candidates who scored high; midrange and passed; midrange and failed; and scored low.

For three of the four candidates, the average score was lower than the original score and for one the score was higher. The candidate who scored midrange and passed was failed by the 15 raters. The original examiners passed 2 of the 4 while the 15 raters passed 1 of the 4. The intraclass reliability coefficients for the raters for two of the candidates were acceptable, .93 and .86 and for two of the candidates the reliability was marginal to unacceptable, .71 and .67.

Two of the raters provided extreme ratings. One rater would have passed all 4 candidates, and one would have failed all four candidates. In summary, there was very little convergence in the licensing decisions between the original examiners and the raters. The candidate's success depends to a large degree upon the pairing of the

examiners. The raters appear to have difficulty in achieving rating consistency when a candidate's performance is near the cut score. The validity of the pass/fail decisions is called into question when the raters evoke radically different decisions, even when they hear the same responses.

CONCLUSIONS

The oral examination does not fully meet psychometric standards required for developing examinations that use judgments of raters making pass/fail decisions. The outcome measure of acceptable internal reliability does not mean that the pass/fail decision is valid. Validity is inferred only if all the standards are embedded in examination development, administration, and scoring. The inability to offer a standardized examination process, calibrate examiners to the same standards, and provide information necessary to make predictions about candidates' behavior indicate that continued use of the oral examination is a questionable practice.

The rationale for this paper from a psychometric point of view is that the oral examination does not provide candidates with a standardized examination so that they receive the same examination experience. It is nonstandardized because of the differences in the examiners' interpretation of the concept of minimum competence when they apply the rating scale. Intraclass correlation coefficients have been used to provide an estimate of the reliability of the examinations. The coefficients have been quite high, indicating that the examiners were consistent. Also, data were analyzed to assess the degree of agreement of pass/fail decisions. There was a fairly high degree of agreement here as well. However, while the internal reliability of the examination was acceptable, the reliability of the judgments about candidates' ability to practice is not acceptable, given the variability of the examiners' pass/fail decisions across candidates within the same administration. The variability of examiners' evaluations and a fairly high failure rate lead to a high rate of false negative errors in the examination process that calls into question the validity of the oral examination.

There may not be a remedy for questionable validity because of the complexity of the examination process and the nature of the constructs being assessed. An important question in the science of measuring individuals is whether the construct of interest actually can be measured with a high degree of confidence. The construct that the oral examination is attempting to measure is the ability of candidates to independently practice at a minimal level of competence—a very difficult construct to operationalize.

Difficulties with the administration and scoring of the examination are not consistent with the standardization necessary to provide each candidate with the same examination experience. The examination is not standardized because of the differences in the examiners and their interpretation of the concept of minimum competence by examiners who provide judgments about the candidates' responses. Different examiners produce vastly different judgments that are not reliable across examination cycles or examination sites.